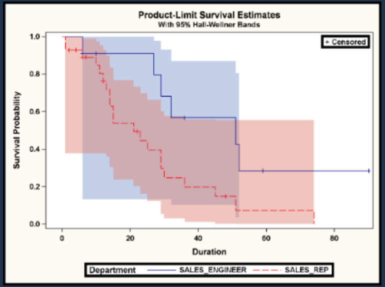
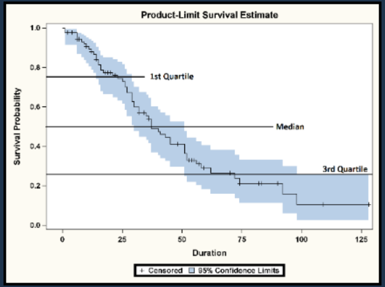




Data Science in Action #1

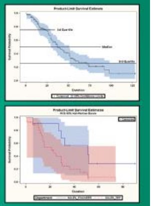
Performing Headcount Survival Analysis for Employee Retention

Gerhard Svolba
Data Scientist, SAS Austria

Data Science Applications and Case Studies

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention



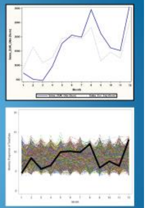
Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

sas

Data Science in Action: #5

Checking the Alignment with Predefined Pattern



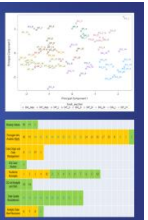
Which customers show a behavior that is far from what you expected?

Chi2 independency test
Benford's law
Time Series Similarity

sas

Data Science in Action: #7

Topic Search Documents and Clustering



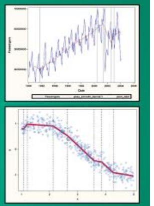
Can I automatically find clusters of documents with similar content?

Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

sas

Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data



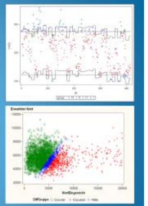
Can events and changes in the course over time be automatically detected?

Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

sas

Data Science in Action: #6

Proving a reference value that considers all available co-information



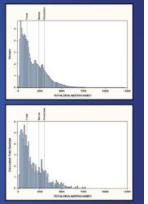
Can analytics help me to reduce the "Yes, but ... " sentences in my business discussions?

Linear Regression
Decision Trees
Time Series Analysis

sas

Data Science in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution



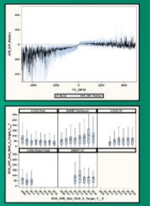
When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?

Monte Carlo Simulations
Mathematical Programming

sas

Data Science in Action: #3

Explaining Forecast Errors and Deviations



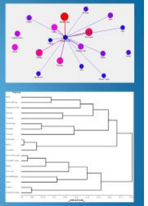
Do the demand planners really improve forecast accuracy with their manual overwrites?

Linear Regression
Quantile Regression
Descriptive Statistics

sas

Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods



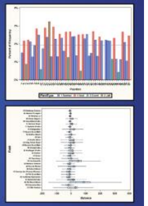
Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

Unsupervised machine learning methods:
association analysis
variable clustering

sas

Data Science in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game



How can you simulate complex environments to get insight in the most frequent processes?

Monte Carlo Simulations

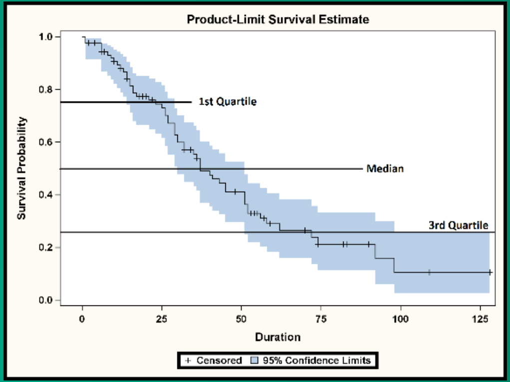
sas

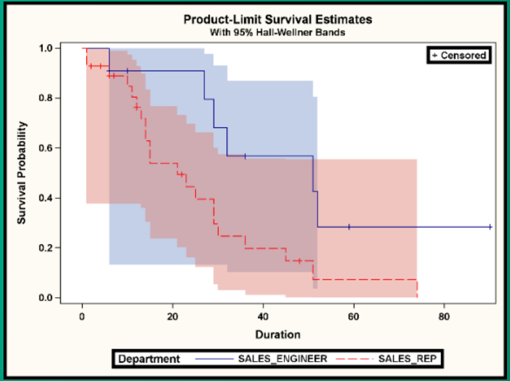
sas

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?






```
Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining
```

Example from the„Human Resources“ Area

- Retention time of employees in a company
- Data Collection: 01/2009 to 12/2016, first employee in 2004
- By department: Marketing, Admin, Sales, TechSupport, Sales Engineer

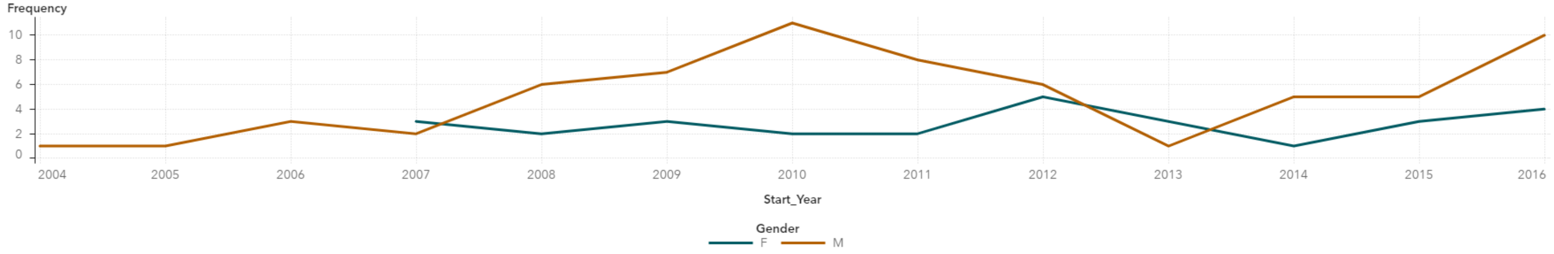
EmpNo	FirstName	Department	Gender	Start	End	Status	Duration
1021	Mary	MARKETING	F	01JUL2009	01AUG2012	0	37
1022	Frank	SALES_REP	M	01JUL2009	01JUN2010	0	11
1023	Alan	SALES_ENGINEER	M	01JUL2009	.	1	90
1024	Frencesca	ADMINSTRATION	F	01AUG2009	01FEB2012	0	30
1025	Karl	SALES_ENGINEER	M	01AUG2009	01DEC2013	0	52
1026	Hana	ADMINSTRATION	F	01AUG2009	01APR2010	0	8
1027	Brian	SALES_REP	M	01NOV2009	01NOV2010	0	12
1028	Pawel	SALES_REP	M	01NOV2009	01APR2012	0	29
1029	Alessandro	TECH_SUPPORT	M	01FEB2010	.	0	83

Performing Descriptive Analyses and Creating Dashboards

Durschnittliche Verweildauer nach Kategorien

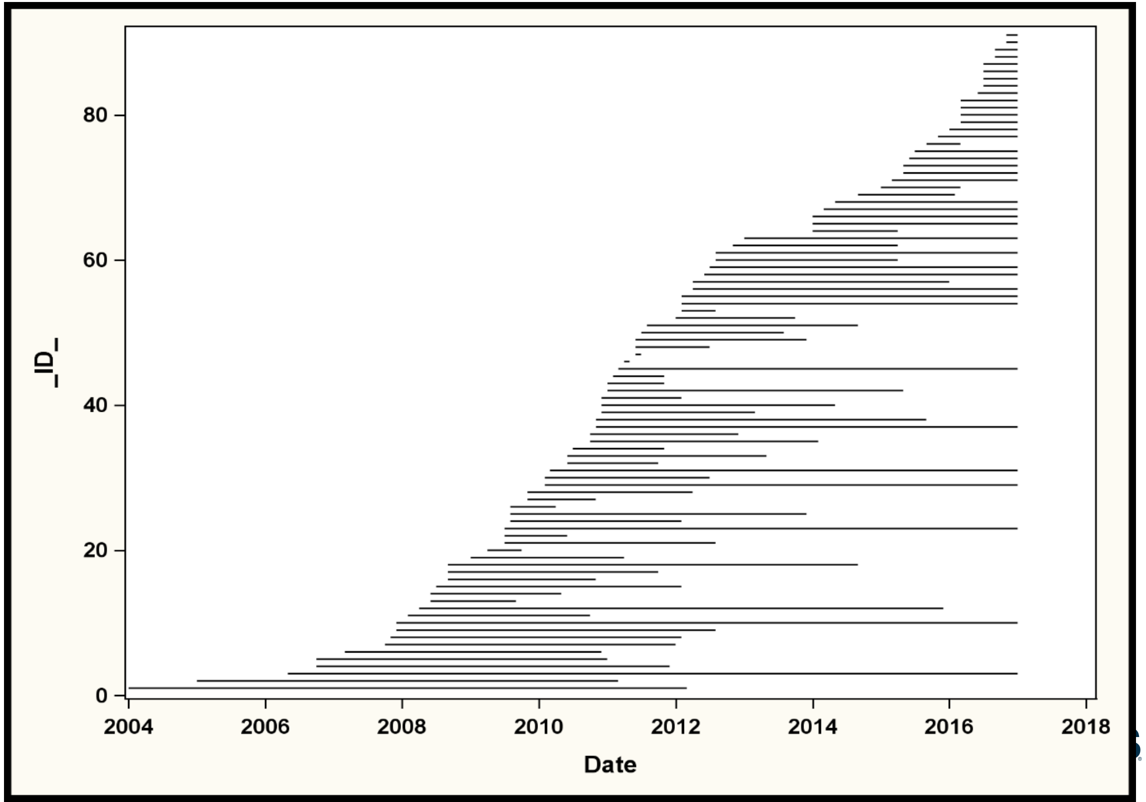
Department	TechKnowHow	Gender	Resigned	Frequency	Duration
TECH_SUPPORT	NO	M	0	2	13
			1	6	29
	YES	F	0	5	35
			1	1	37
		M	0	8	52
			1	8	32
SALES_REP	NO	F	1	3	15
		M	0	7	14
			1	18	24
SALES_ENGINEER	YES	M	0	5	40
			1	6	33
MARKETING	NO	F	0	3	26
			1	1	37
		M	0	1	57
			1	3	83
ADMINSTRATION	NO	F	0	4	35
			1	8	40
		M	0	2	72

Frequency of Start_Year grouped by Gender



We do not have an event date for all employees (luckily ☺)!

- Observe Careers per Employee
  - Different length
  - „Left company“ or „censored“



Business Questions

- What is the average retention period for employees in the company?
  - How can the important fact that the employment end date is known only for those who already left the company, be adequately considered in the analysis?
- How can the retention period be visualized and compared between different subgroups?
- Are there influential factors for the length of the retention period?
- How can these factors be ranked by magnitude of their influence?
- Can the expected survival period for an employee be predicted?

How can we deal with missing endpoints?
→ Kaplan-Meier Analysis

Sales-Engineer Department

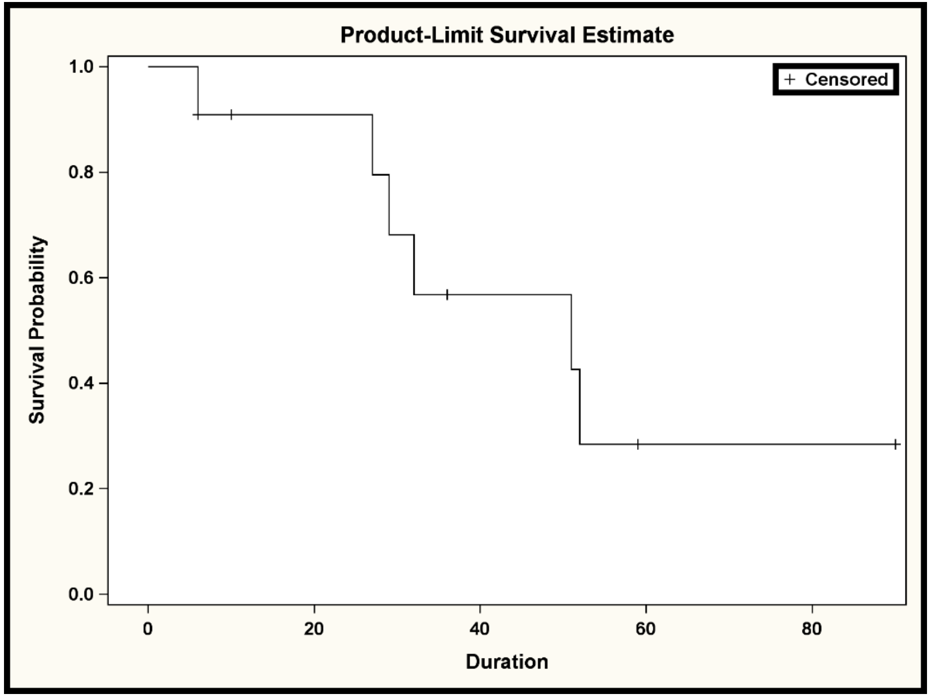
Duration	Left	Resigned	Censored	Survival	Comment
0	11			1,000	Start of Observation
6	10	1	0	0,909	John resigns
6	9	0	1		Brady is censored from the analysis
10	8	0	1		Lucas is censored from the analysis
27	7	1	0	0,795	Rainer resigns
29	6	1	0	0,682	Vincenz resigns
32	5	1	0	0,568	George resigns
36	4	0	1		Mark is censored from the analysis
51	3	1	0	0,426	Viktor resigns
52	2	1	0	0,284	Karl resigns
59	1	0	1		Eugene is censored from the analysis
90	0	0	1	0,284	Alan is censored from the analysis

Kaplan-Meier Analysis allows you to estimate the median and average rentention period

```
proc lifetest data=employees ;
 time Duration*Status(1);
 where Department='SALES_ENGINEER';
run;
```

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	32.0000	.
50	51.0000	LOGLOG	27.0000	.
25	29.0000	LOGLOG	6.0000	51.0000

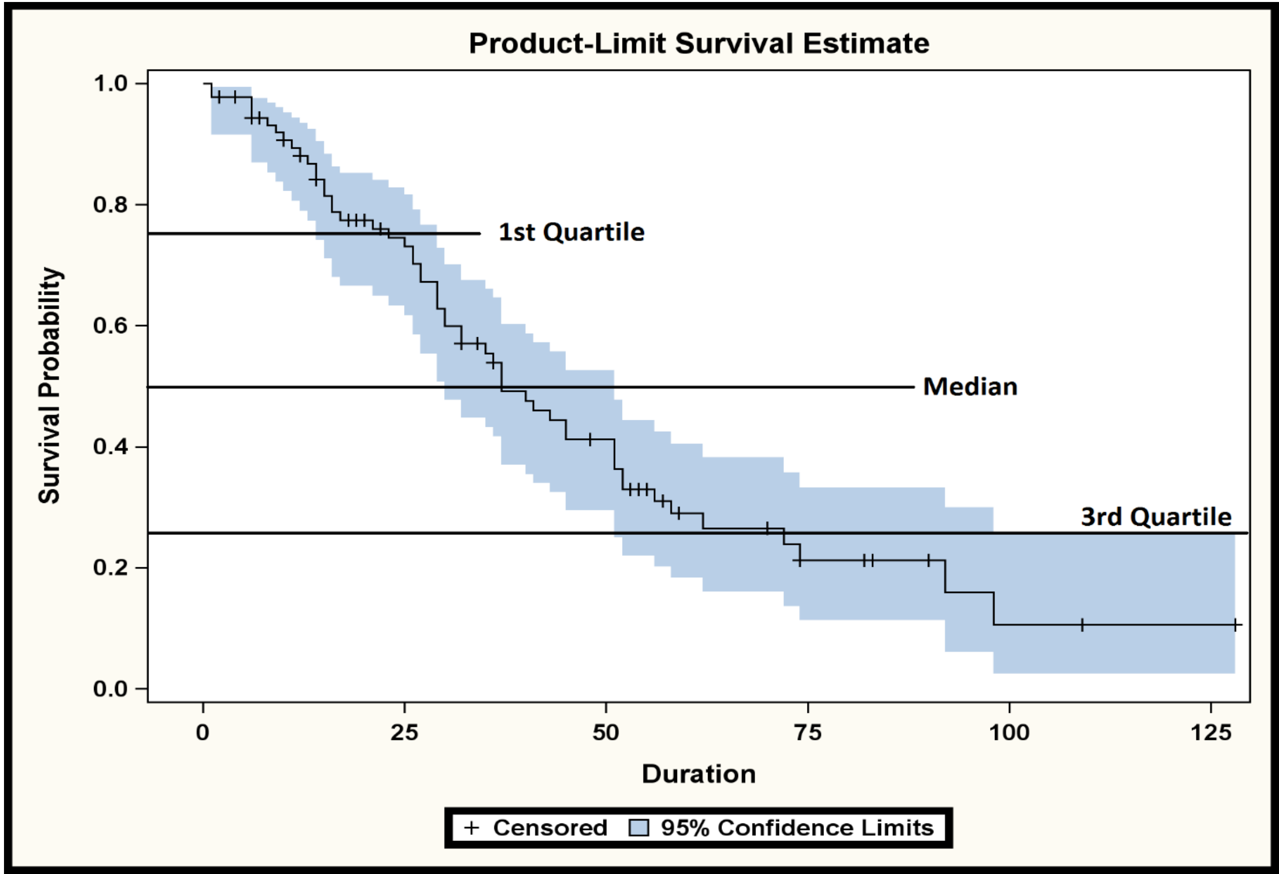
Mean	Standard Error
39.9489	5.2333



Looking at the retention period for all employees. Interpretating the Survival Kurve

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	72.000	LOGLOG	51.00	.
50	37.000	LOGLOG	30.00	51.00
25	23.000	LOGLOG	14.00	29.00

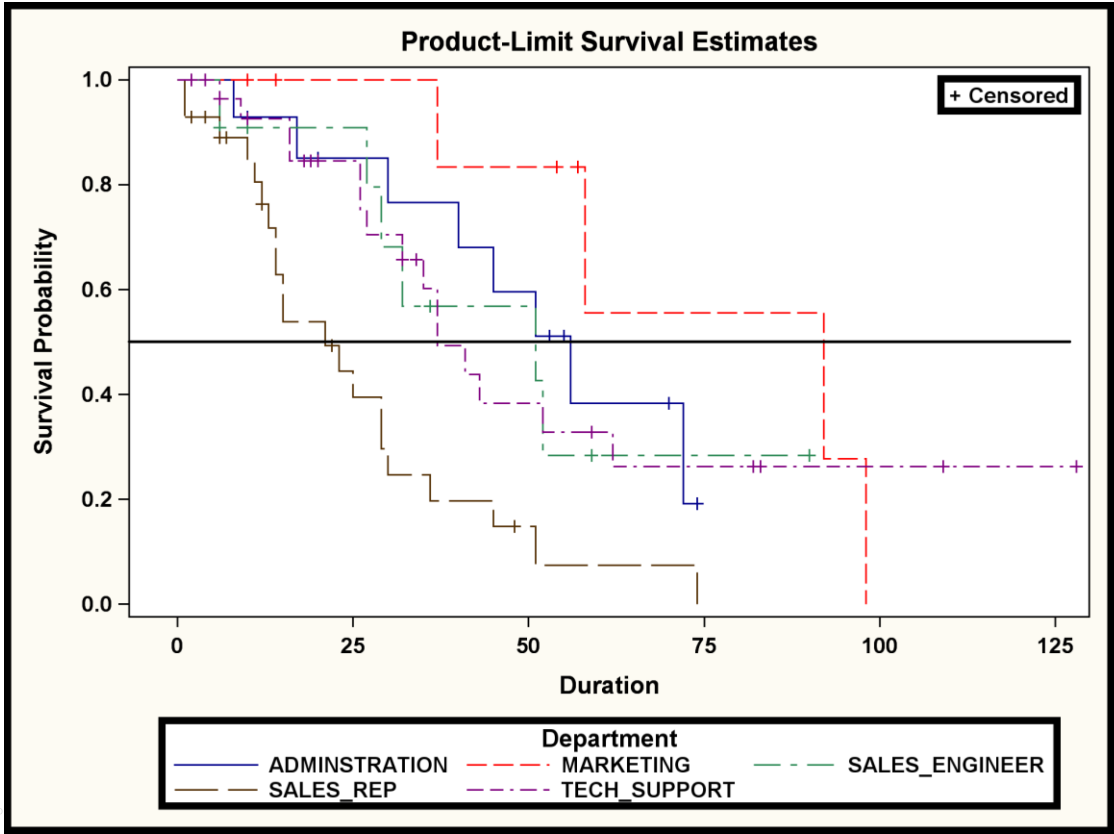
Mean	Standard Error
46.757	3.813



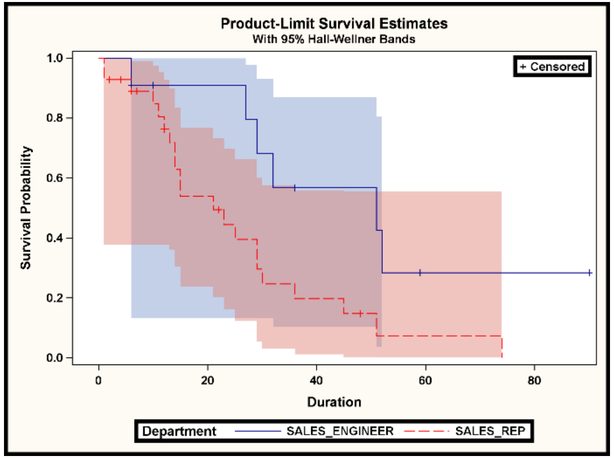
Can we compare the analysis between departments?

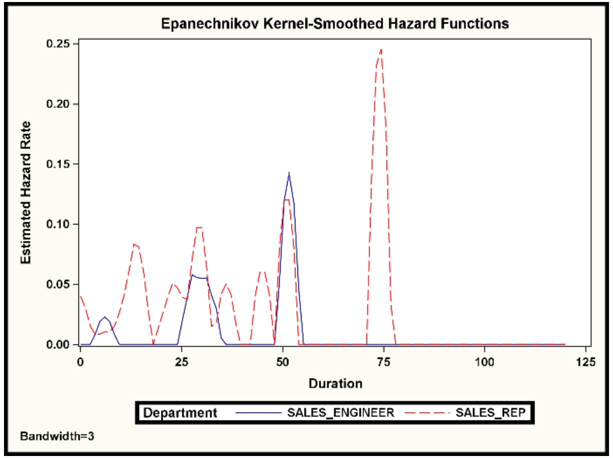
Running the analysis per Department

```
PROC LIFETEST DATA=employees;
 TIME Duration*Status(1);
 STRATA department;
RUN;
```



Comparing selected departments and studying the hazard curve per department





Kaplan Meier Methods and Cox Proportional Hazards Regression:
Sales engineers have a better survival time than sales representatives.

.

Studying the Hazard Curves: There is high risk to lose your sales engineers after 26 and after 50 months.

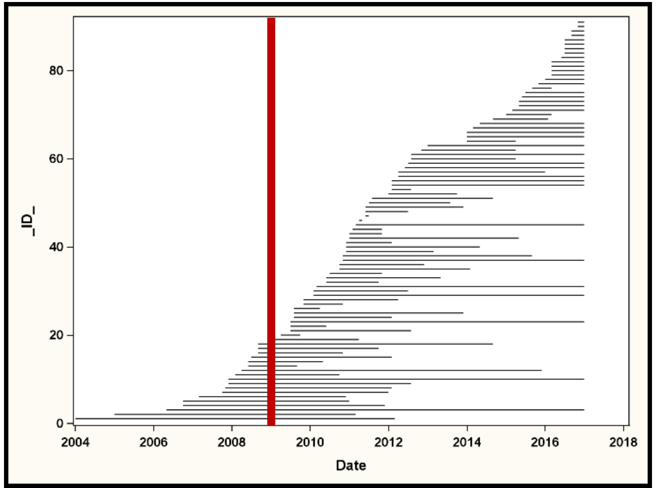
In the „good old times“
everything has been better!
Employees were more loyal and stayed longer.

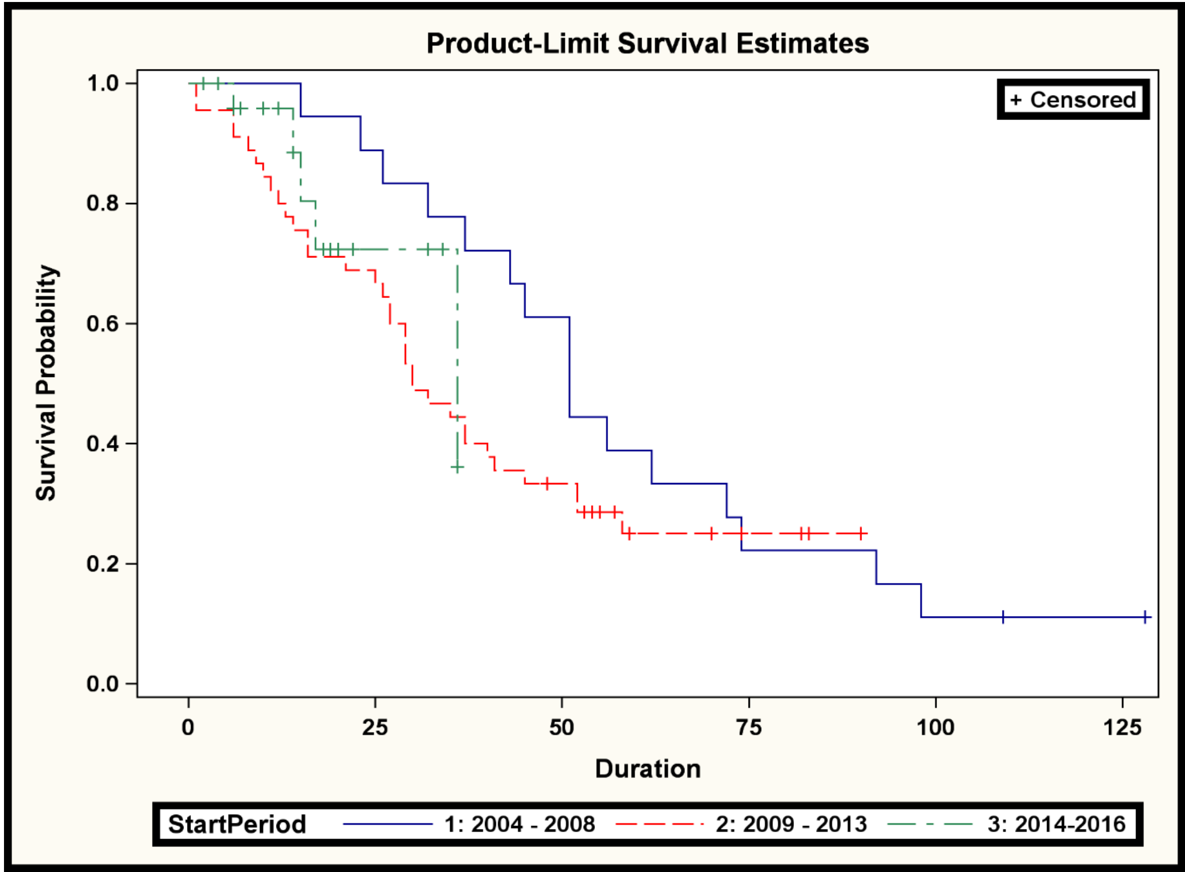
Really?

Consider how your data have been collected!

Stratifying the analysis per „Start Period“

- Data Collection: 01/2009 to 12/2016, first employee in 2004
- „Pre-Selection“ of the data





What are the most influential factors for employee retention?

Can we perform predictive modeling on censored data?

How long will Gerhard still stay in our company?

Given certain risk factors, what is the expected survival in 6 months and the probability to resign within the next 6 months.

EmpNo	Department	Gender	TechKnowH...	_T_	EM_SURVFCST	EM_SURVEVENT	T_FCST
1003	TECH_SUPPORT	M	YES	128	0.240	0.000	134
1010	TECH_SUPPORT	M	YES	109	0.240	0.011	115
1023	SALES_ENGINEER	M	YES	90	0.108	0.313	96
1029	TECH_SUPPORT	M	YES	83	0.386	0.133	89
1031	TECH_SUPPORT	F	YES	82	0.177	0.219	88
1037	ADMINSTRATION	M	NO	74	0.471	0.066	80
1045	ADMINSTRATION	M	NO	70	0.494	0.053	76
1054	TECH_SUPPORT	F	YES	59	0.316	0.102	65
1055	SALES_ENGINEER	M	YES	59	0.313	0.103	65

Use the Cox-Proportional-Hazard Regression to perform regression analysis on censored data

```
PROC PHREG DATA=Employees;
 CLASS department gender TechKnowHow
       /PARAM=effect REF=first;
 MODEL Duration*Status(1)=
         department gender TechKnowHow
       /SELECTION=stepwise;
RUN;
```

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Department	MARKETING	1	-1.15513	0.47794	5.8414	0.0157	0.606
Department	SALES_ENGINEER	1	0.82336	0.52244	2.4838	0.1150	4.380
Department	SALES_REP	1	0.62976	0.29224	4.6436	0.0312	3.609
Department	TECH_SUPPORT	1	0.35572	0.29940	1.4117	0.2348	2.744
TechKnowHow	YES	1	-0.63474	0.27370	5.3781	0.0204	0.281

Watch my webinar
Interpreting Machine Learning Models,
to see how to display the value for the reference category!

The model allows you to output the predicted Survival for 24 months in the future for the existing employees

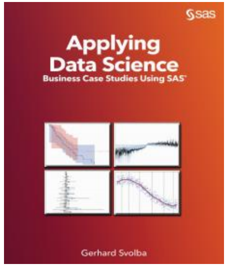
	EmpNo	FirstName	Department	TechKnowHow	Gender	Start	End	S_Duration_4
1	1088	Simone	TECH_SUPPORT	YES	F	2016-09-01	.	0.8264229844
2	1091	Guido	SALES_REP	NO	M	2016-11-01	.	0.5661954848
3	1087	Serge	SALES_REP	NO	M	2016-07-01	.	0.5661954848
4	1080	Nina	TECH_SUPPORT	YES	F	2016-03-01	.	0.8264229844
5	1059	Verena	MARKETING	NO	F	2012-07-01	.	0.8593533102
6	1074	Manuel	TECH_SUPPORT	NO	M	2015-06-01	.	0.6361124813
7	1084	Jean	TECH_SUPPORT	NO	M	2016-07-01	.	0.6361124813
8	1023	Alan	SALES_ENGINEER	YES	M	2009-07-01	.	0.8188481424
9	1075	Olivier	TECH_SUPPORT	YES	M	2015-07-01	.	0.9049068956
10	1031	Lisa	TECH_SUPPORT	YES	F	2010-03-01	.	0.8264229844
11	1003	Jim	TECH_SUPPORT	YES	M	2006-05-01	.	0.9049068956
12	1079	Francesca	ADMINSTRATION	NO	F	2016-03-01	.	0.8303156037
13	1056	Bob	MARKETING	NO	M	2012-04-01	.	0.9236297793
14	1072	Bettina	TECH_SUPPORT	YES	F	2015-05-01	.	0.8264229844
15	1085	Joshua	TECH_SUPPORT	YES	M	2016-07-01	.	0.9049068956
16	1067	Joseph	TECH_SUPPORT	YES	M	2014-03-01	.	0.9049068956
17	1068	Timon	TECH_SUPPORT	YES	M	2014-05-01	.	0.9049068956
18	1081	Anja	MARKETING	NO	F	2016-03-01	.	0.8593533102
19	1045	Malcolm	ADMINSTRATION	NO	M	2011-03-01	.	0.9071383626
20	1010	Paul	TECH_SUPPORT	YES	M	2007-12-01	.	0.9049068956
21	1029	Alessandro	TECH_SUPPORT	YES	M	2010-02-01	.	0.9049068956
22	1061	Alice	ADMINSTRATION	NO	F	2012-08-01	.	0.8303156037

Conclusion

- Data Science methods provide insight where simple descriptive methods fail: „Censored Data“.
- You can study the findings between subgroups and compare them.
- Cox-Prop.Hazard Regression allows to perform regression analysis on censored data.
- Make sure that you understand how your data is collected!

Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis



Get access to more content:

SAS DACH @Youtube:	https://www.youtube.com/user/SASsoftwareGermany
Blogs on LinkedIn:	https://www.linkedin.com/in/gerhardsvolba/
Twitter:	https://twitter.com/gsvolba
Content on Github:	https://github.com/gerhard1050
Books @SAS-Press:	https://support.sas.com/svolba